# Sir Han Sloane's Information Architecture

Lessons Learned From A Data Driven Research

Deborah Leem, Julianne Nyhan, Antonis Bikakis

There are varying levels of complexities that need to be tackled and addressed when humanities sources are conceptualised and processed as data. This paper seeks to examine some of the challenges and future possibilities of treating early-modern collection catalogues as data through a case study: automatic extraction of place and person names from a TEI-XML document.

## Background

Sir Hans Sloane (1660-1753) bequeathed his immense and diverse collection of objects[1] to the British Nation upon his death. This collection became the foundation of the UK's three national memory institutions: the British Museum, the Natural History Museum, and the British Library.

Sloane's own manuscript catalogues of his collections are fundamental to unlocking his intellectual legacies. As part of a collaborative project between UCL and the British Museum, the Leverhulme-funded 'Enlightenment Architectures: Sir Hans Sloane's catalogues of his collections'[2] (2016–19) arranged for the capture, via double keying[3], of 5 volumes of Sloane's manuscript catalogues as digital data. They were subsequently encoded by the Enlightenment Architectures (EA) team in line with a project-modified schema of the Guidelines of the Text Encoding Initiative (TEI).

Building upon the EA project, my PhD will conduct a data-driven analysis of Sir Hans Sloane's catalogues of his collections. One of the aims of my research is to leverage the mark-up of these catalogues to derive new computational outputs that are amenable to subsequent data analysis. Furthermore, approaching the research from a 'Collections as Data'[4] perspective, my research goal is to provide a critical analysis of data modelling and its potential impact on future data-driven research in the context of Early Modern documents. In this paper I will present a case study from my ongoing research that offers an exemplary way of exploring these issues.

[1] Sloane amassed a diverse collection and upon his death prints, drawings, books, manuscripts, herbarium, antiquities along with other treasures were offered to the British nation.

[2] Project information and digital editions available at https://reconstructingsloane.org/. ("Reconstructing Sloane," n.d.)

[3] This process was outsourced to AEL Data (https://aeldata.com/)

[4] See e.g. https://collectionsasdata.github.io/

## Case Study

This case study will explore some of the situated difficulties that can arise during the automatic extraction of place and person names from catalogue records for network analysis. The focus of this case study is the manuscript catalogue volume titled ‘Miscellanea’ that is formed of seven different catalogues. The Python programming language and Pandas, a data analysis framework, have been instrumental in extracting targeted data from the TEI-XML encoded representation of the catalogue.

In this case study, the extraction routine aims to retrieve all the place and person names that were encoded in the numerous catalogue entries of which the Sloane’s catalogues are comprised, along with their catalogue numbers. This is necessary in order to establish a canonical spelling and id, to which the variant place and person name forms (a common feature of early-modern texts) will all point. Moreover, it is expected that an added bonus of this work will be the way that it may enrich the existing authority files, which are currently uneven for the early-modern period.[5]

As Winters points out, humanities data is ‘messy and unpredictable.’ (Winters, 2017) Heterogeneity is one of the features that comes up when processing data and this brings challenges. It would be reasonable to have expected that the TEI mark-up that had already been applied to the text would have allowed us to harness this heterogeneity. Yet, in order to produce a complete list of person and place names, it was necessary to implement multiple iterations of script refinement and revision, due to the complexities of the TEI-XML schema.

Anomalies and inconsistencies in the catalogue are some of the contributing factors. In Miscellanea, there are 4569 unique catalogue numbers and 39 duplicate numbers assigned to the respective catalogue entries. Most of these ‘duplicates’ are due to the fact that when Sloane or a later amanuensis ran out of space while recording some of the catalogue entries, they continued on the verso side of the pages with the same catalogue number.[6] There are also notes and additions belonging to the existing records which may contain important information pertaining to objects in the catalogue. Moreover, some of these notes and additions are connected to a catalogue entry through the use of an asterisk, ‘*libra ponda*’ (hashtag) symbol or a plus sign, in effect, creating both implicit and explicit cross references between catalogues entries. Further, information captured and encoded within the catalogue number element required data clean up. This is largely due to the catalogue numeric numbers also containing dashes, dots, question marks, underscores or asterisk. These anomalies were faithfully captured and marked-up by the EA team as their focus was primarily on the act of modelling and presenting historically sensitive readings of the catalogues (Ortolja-Baird et al., 2019). TEI mark-up enhanced the catalogue data in the ways

[5] There are projects like Archaeology of Reading and Circulation of Knowledge and Learned Practices in the 17th Century Dutch Republic (CKCC) that illustrate difficulties with dealing with different variations of place and person names. Automatic detection of named entities especially in the Early Modern documents is still problematic and requires manual normalisation to a great extent.

[6] Some of these numbers that are used twice in the catalogue appear to be mistakes made by Sloane or his amanuensis and it only came to our attention during the data cleaning process.

that would allow more sophisticated analysis. However, difficulties with tackling the complexities of the data remain.

## Conclusion

The intended use and target audience greatly affect the thinking behind the data modelling. Thus, the decisions made in the process of transforming historical analogue texts into digital data have huge impacts and implications on the reusability of the data. By approaching the data models and standards like TEI and CIDOC-CRM[7] from a Collections as Data perspective, this paper shares experiences and lessons learned from applying computational methods to create reusable datasets. Furthermore, by providing concrete examples of the possibilities and challenges of the approaches taken from the case study, this paper examines how amenable the TEI mark-up of the Early Modern catalogues is to computational methods.

Reflections from the case study have a role in addressing challenges, limitations and potential benefits of contributing to the Collections as Data movement or other data modelling projects. This also adds value to data driven humanities research by demonstrating how new knowledge and insights have risen from the use of digital methods in the context of Early Modern documents.

## Bibliography

Always Already Computational [WWW Document], n.d. . Always Already Comput. - Collect. Data. URL https://collectionsasdata.github.io/ (accessed 3.3.20).

Ortolja-Baird, A., Pickering, V., Nyhan, J., Sloan, K., Fleming, M., 2019. Digital Humanities in the Memory Institution: The Challenges of Encoding Sir Hans Sloane's Early Modern Catalogues of His Collections. Open Libr. Humanit. 5, 44. https://doi.org/10.16995/olh.409

Reconstructing Sloane [WWW Document], n.d. . Reconstr. Sloane. URL https://reconstructingsloane.org/ (accessed 3.3.20).

Winters, J., 2017. Tackling complexity in humanities big data: from parliamentary proceedings to the archived web., in: Big and Rich Data in English Corpus Linguistics: Methods and Variations, Studies in Variation, Contacts and Change in English. VARIENG, Helsinki.

[7] Another case study from my research focuses on linking TEI and external ontologies using CIDOC-CRM.